Popularity Prediction on Turkish News Stories

Eran Toker
Hacettepe University
Ankara, Turkey
Email: erantoker@gmail.com

Fatih Güler
Hacettepe University
Ankara, Turkey
Email: ffguler@gmail.com

Yiğit Sever
Hacettepe University
Ankara, Turkey
Email: yigitsever94@gmail.com

***Abstract*—News agencies publish numerous articles on their websites but only a small portion of those stories get readers' attention and become popular. Although popularity can be measured through the clicks a certain article get, users will also 'vote' on the popularity of stories through their social media shares of articles and this may give opinions about popularity of news story. In this project, we are trying the find out if the news article published by Milliyet.com.tr will be shared in EksiSozluk.com, a popular Turkish discussion forum.**

I. Introduction

The advancement and steady growth in social media, mainly blogosphere and streaming sites, allowed new types of data to become available which can be mined for valuable knowledge. To give an example; online discussions can be used to predict sales ranks of books. [1] Copious amounts of posts on social media are posted as a response to events that users read from news articles, as a result, investigating events and their social impact that is reflected in social media has become an important task for media analysts.

Our purpose in this work is predicting the popularity of a news article before it goes live. Predicting popularity and traffic of an article might help with determining the desirability of an article, whether or not it is worth to publish and advertise on it. [2] If news agents knew which articles can or will be popular, they can spend their resources on potential candidates. This concept is also important in the sense of serving a better user experience; news sources can offer articles which might be more interesting to readers. If online news agencies can gain more users, they can earn more money from ad revenue. [3]

Our aim in this paper is to predict popularity of news articles before publication. Popularity is used in the sense that whether or not the article will be shared on EksiSozluk.com which is the most popular social media establishment in Turkey, with Turkish origins.

This work makes several contributions. First, it explores the dynamics of stories in Milliyet.com.tr. Second, it introduces the problem of predicting the popularity of a news article. Third, it provides a set of textual and semantic features that can be used to predict if a news article will be shared or not in EksiSozluk.com before the said article is published. Fourth, it provides an evaluation of the introduced features. Fifth, an error analysis identifies possible causes for classification failure.

II. Related Work

Most of the popularity prediction works based their approach on early comments/likes/retweets. [4] The implication is, the content actually needs to be published, losing valuable aspects that will be gained from a prior analysis. In a research conducted in 2009 [5], they have used a 5 feature set for finding popularity; surface, cumulative, textual, semantic and real world. Surface feature consisted of images present in the article or how many authors have contributed to the article. Cumulative feature looked for how many articles were published at the same hour. Real world feature analyzed whether or not the weather was nice or cloudy. Subsequently, we used *textual* and *semantic* features since those are the only features that are related to text mining.

III. Data and Features

A. Exploring News

The data we used in the project consists of 449 news stories published at Milliyet.com.tr between February 2016 to August 2016. Among those 449 news stories, 60 of them are tagged as 'unpopular' by us since they were not shared in Eksisozluk.com. The reason for relatively small sample size is that there were no set methods or data sets of Turkish news articles that are linked to threads opened at Eksisozluk.com, as a result, the dataset was put together by us via hand. Another challenge was the fact that Eksisozluk.com is not a social media that users share content with links and have discussion based on them (i.e. Twitter, Reddit) but a discussion forum with topics and thread titles decided by users. Therefore, an automated method would not be plausible.

The selected news articles were then parsed and their content extracted. News articles come with a heading, a subheading and the actual text content which we appended together and handled as news body.

Our first research question is, what are the dynamics of user generated comments on news articles? According to Tsagkias et al. [5], there were 5 features with two of them about text mining that we used. Textual feature means getting terms at top 200 frequency in our corpus. Terms were ranked by their log-likelihood scores for all stories. Semantic feature is the count of people, location and organization mentioned in the articles.

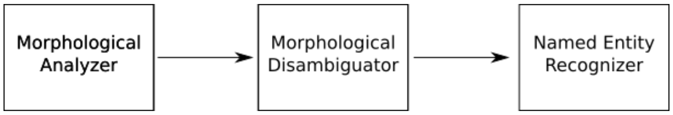

Fig. 1. Pipeline Schema.

B. Feature Engineering

For textual feature, we first found most common 200 terms in all stories. We skipped stop words in Turkish and used Resha Turkish Stemmer [6] for stemming operations. As we initially predicted; ”Basbakan”, ”Cumhurbaskani”, ”Futbol”, ”Polis”, ”Asker”, ”Sehit”, ”Patlama”, ”Teror” tokens were some of the most common terms as those were related to common occurrences in Turkey, 2016.

For semantic feature, first we tried using ITU Natural Language Processing Toolkit. With this tool, we were able to find Named Entities of each word in the articles. According to pipeline we used the schema represented in Figure 1. Thus we were able to get person, organization and location counts for each news article. However, this tool was unfeasibly slow; a query for one single word took more than a minute at times. Therefore, we switched to DBPedia-Spotlight [7]. This tool offered in-house querying instead of API querying and we could get person, organization and location results. After that we used Random Forest Algorithm [8]. We had 4 numerical data as in Term Frequency of popular terms, Person Count, Organization Count and Location Count as well as one label which is popular or not. Consequently, we reimplemented GitHub user ’ironmanMa’s algorithm for our purpose.

IV. Experiment Setup

As mentioned beforehand, EksiSozluk.com does not provide an API, so we manually found news stories shared in EksiSozluk.com. After creating a heap of stories that have been talked about in EksiSozluk.com, we extracted the news stories from webpages and categorized them by month. Then, most common 200 terms were found and written to ’MostCommonTerms’ file. Later, we processed input file for textual features and wrote to ’TermFrequencies’ file. Using DBpedia Spotlight, threads wrote to output file ’SemanticVariables’ fir each month. Via 3 output files, we have created ’TreeData’ files for each test and training stories and wrote to ’TreeData’ and ’TestTreeData’ file. Using tree data files we used random forest implementation for results. Then we have calculated F1 scores and accuracy.

V. Results

The results looked promising as shown in Table II, approaching results achieved by Bandari et al. [3] with precision score of 0.8.

As mentioned before, we opted to convert our dataset to lowercase for the experiments. However, this lead DBpedia Spotlight tool to miss person information, while not having an impact on location or organization information. To eliminate this error, we reverted our dataset back to sentence case (unaltered case) and ran the tests again. Results are shown in Table III and IV and are improved from lowercase dataset tests by a noticeable margin.

TABLE I
Test Results for Lowercase dataset

True Positive	33
False Positive	8
False Negative	15

TABLE II
Precision, Recall and F1 Scores for Lowercase dataset

Precision	0.80
Recall	0.69
F1 Score	0.74

TABLE III
Test Results for Sentence Case dataset

True Positive	36
False Positive	7
False Negative	12

TABLE IV
Precision, Recall and F1 Scores for Sentence dataset

Precision	0.83
Recall	0.75
F1 Score	0.79

VI. Conclusion

In this work, we aimed to predict the popularity of news articles prior to their publication. We have created a dataset and picked our features regarding earlier works in the literature. After the evaluation steps, we have looked into the misclassified stories. Most of the stories that have been labeled as ’popular’ by our implementation but do not have a EksiSozluk.com thread associated to it are actually popular but just not by our parameters. These false positives did have discussions about them, but on threads of a popular person mentioned in the article. We did not have an autonomous method of finding these links. Another case of misclassified stories included a lot of famous people or organizations but the stories themselves were not interesting or worth talking about (i.e. clickbait).

References

[1] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, “The Predictive Power of Online Chatter,” in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ser. KDD ’05. New York, NY, USA: ACM, 2005, pp. 78–87. [Online]. Available: http://doi.acm.org/10.1145/1081870.1081883

[2] D. Phukan and A. K. Singha, “Feasibility analysis for popularity prediction of stack exchange posts based on its initial content,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 1397–1402.

[3] R. Bandari, S. Asur, and B. A. Huberman, “The Pulse of News in Social Media: Forecasting Popularity,” *arXiv:1202.0332 [physics]*, Feb. 2012, arXiv: 1202.0332. [Online]. Available: http://arxiv.org/abs/1202.0332

[4] G. Szabo and B. A. Huberman, “Predicting the Popularity of Online Content,” *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010. [Online]. Available: http://doi.acm.org/10.1145/1787234.1787254

[5] M. Tsagkias, W. Weerkamp, and M. de Rijke, “Predicting the Volume of Comments on Online News Stories,” in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM ’09. New York, NY, USA: ACM, 2009, pp. 1765–1768. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646225

[6] H. R. Zafer, “hrzafer/resha-turkish-stemmer.” [Online]. Available: https://github.com/hrzafer/resha-turkish-stemmer

[7] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, “Improving efficiency and accuracy in multilingual entity extraction,” in *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[8] M. Arafath, “ironmanMA/Random-Forest.” [Online]. Available: https://github.com/ironmanMA/Random-Forest